

## Chapter 4

# Association between Two or More Variables

Very frequently social scientists want to determine the strength of the association of two or more variables. For example, one might want to know if greater population size is associated with higher crime rates or whether there are any differences between numbers employed by sex and race. For categorical data such as sex, race, occupation, and place of birth, tables, called *contingency tables*, that show the counts of persons who simultaneously fall within the various categories of two or more variables are created. The Bureau of the Census reports many tables in this form such as sex by age by race or sex by occupation by region. For continuous data such as population, age, income, and housing the strength of the association can be measured through correlation statistics.

### A. Cross Tabulations

Contingency tables such as that below are quite popular because they are easy to understand and can be used with nominal, ordinal, interval, or ratio data. In such a table it is easy to see the frequency of persons that belong to the categories of both variables. For higher measurement levels, the variables are typically coded into several categories such as less than 18 years, 18 to 64 years, and 65 and older.

One of the most common measures of association for contingency tables is Chi-square. With this statistic we compute the expected frequencies for the cells which would represent the case that there is no relationship among the variables. As the actual numbers depart from the expected values, the larger and more significant Chi-square becomes. The significance level of Chi-square depends on the number of observations and the number of cells in the table and so for census data, which often has very large counts, small deviations from the expected values will be statistically significant. Chi-square also expects at least 5 cases in each cell in order to estimate values reliably.

For this particular table one might expect the marital status of males and females to be about the same. However, the percent of widowed and separated females greatly exceeds that for men.

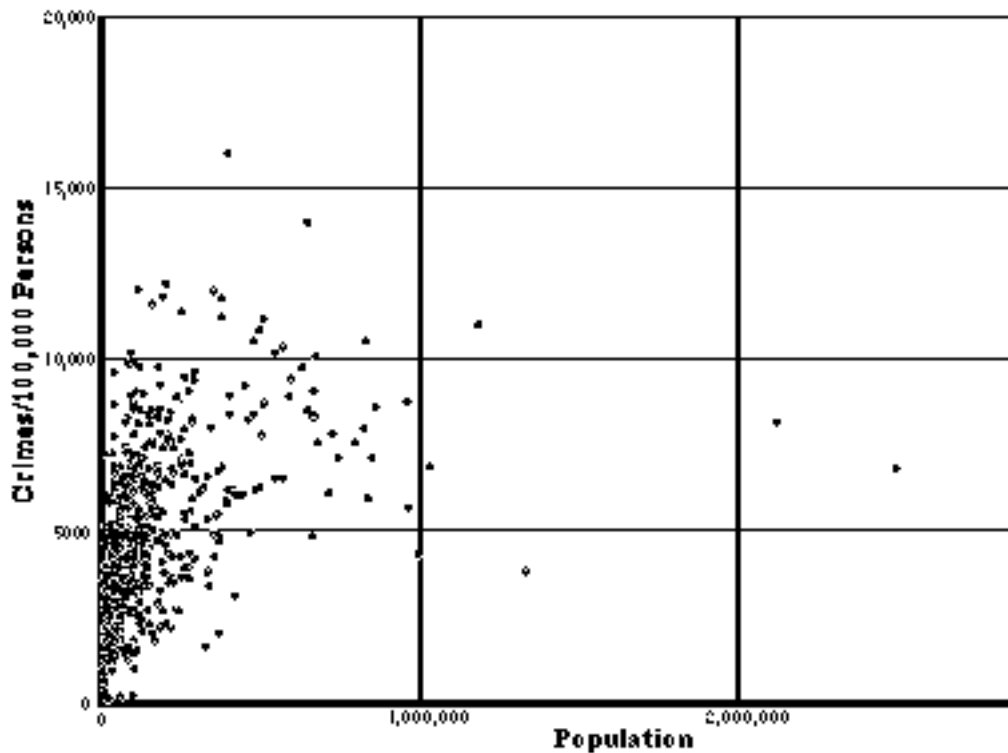
**Table P18 – Sex by Marital Status for Persons >= Age 15**

California, 2000	Male:	Female:	Pct Male:	Pct Female:
Never married	4,343,790	3,500,117	55.4	44.6
Now married:	7,205,642	7,094,229	50.4	49.6
Married, spouse present	6,226,504	6,244,539	49.9	50.1
Married, spouse absent:	979,138	849,690	53.5	46.5
Separated	256,459	386,211	39.9	60.1
Other	722,679	463,479	60.9	39.1
Widowed	278,180	1,179,638	19.1	80.9
Divorced	1,017,057	1,457,510	41.1	58.9

## B. Scattergrams

Scattergrams graphically portray how closely changes in one continuous variable correspond to changes in another. In the example below the population values for the 593 metropolitan counties in the U.S. have been plotted on the x-axis and the corresponding crimes per capita have been plotted on the y-axis.

**Scattergram of Population vs Crimes Per 100,000 Persons**



In this scattergram there does appear to be some association between higher crime rates and larger populations in counties. However, there is quite a bit of variability in this trend—a few cities with large populations have relatively low crime rates and a few small cities have relatively high crime rates. If the relationship was very strong, the points would spread out along a line and if it was very weak, the points would be scattered randomly over the plot. Very strong, almost linear, distributions may be found in physical relationships such as the increase in pressure in a container with an increase in temperature. However, such strong relationships are rare among social data.

## C. Correlation

If a scatter of points does seem to exhibit a non-random trend, then one might choose to measure the strength and the direction of it through the use of correlation statistics. Correlation determines whether a relationship exists between two variables. If an increase in the first variable, x, always brings the same increase in the second variable, y, then the correlation value would be +1.0. If the increase in x always brought the same decrease in the y variable, then the correlation score would be -1.0. If an increase in x brought no regular change in y, then the correlation would be 0. In most calculations of correlation, an approximation of a linear relationship is assumed. However, the relationship could be curvilinear or cyclical, and so one should always examine a scattergram to see if the relationship between two values is non-linear.

There are several types of correlation measures that can be applied to different measurement scales of a variable (i.e. nominal, ordinal, or interval). One of these, the *Pearson product-moment correlation coefficient*, is based on interval-level data and on the concept of deviation from a mean for each of the variables. A statistic, covariance, is the product of the deviations of the observed values from each of their means divided by the number of observations. This mean deviation is divided by the product of the standard deviations of the two variables to get the correlation or:

$$r = \frac{\frac{\Sigma(X - \Sigma X/N) \times (Y - \Sigma Y/N)}{N}}{\text{SQRT} \left[ \frac{\Sigma(X - X)^2}{N} \right] \times \left[ \text{SQRT} \left[ \frac{\Sigma(Y - Y)^2}{N} \right] \right]}$$

The correlation statistic above is for the entire population. If a sample had been selected, the N would have been replaced by n-1.

Computing the Pearson product moment correlation for the crime and population data yields a correlation score of .449, which is only a moderate value. Another statistic, called the *coefficient of determination*, can be calculated to determine the percent of the total variance explained by the correlation between the two variables. The coefficient of determination is simply the square of the "r" or correlation coefficient. In this example, the coefficient of determination is only .202. Thus, about 20% of the variance between population size and crime rate is accounted for by the correlation between these two variables. This would suggest that other variables yet unaccounted for are causing 80% of the crime rate differences between cities..

Since the scatter of points rises steeply and then stretches to the right, a non-linear regression line may fit better than a straight line. Calculating the natural logarithm of the population generates a line that curves to the right. This increases the correlation coefficient to .605 and the coefficient of determination to .367. Thus, a non-linear form of

correlation increases the percent of variance explained to about 37%. Apparently the crime rate does increase with population size, but at a decreasing rate.

Because all 593 metropolitan counties in the U.S. were used to compute the correlation statistic, there is less value to testing its significance. Had a sample of the counties been taken, one could consider the possibility that such a relationship could have occurred by chance. To test the significance of the relationship, one could assume that there is no relationship between population size of counties and the crime rate (null hypothesis) and that the value of  $r$  is due to sampling error. A statistic called the  $t$  statistic is commonly used to test the hypothesis that the correlation value is due to sampling error.

$$t = \frac{|r| \times \text{SQRT}(n-2)}{\text{SQRT}(1 - r^2)} \quad [\text{SQRT} = \text{square root}]$$

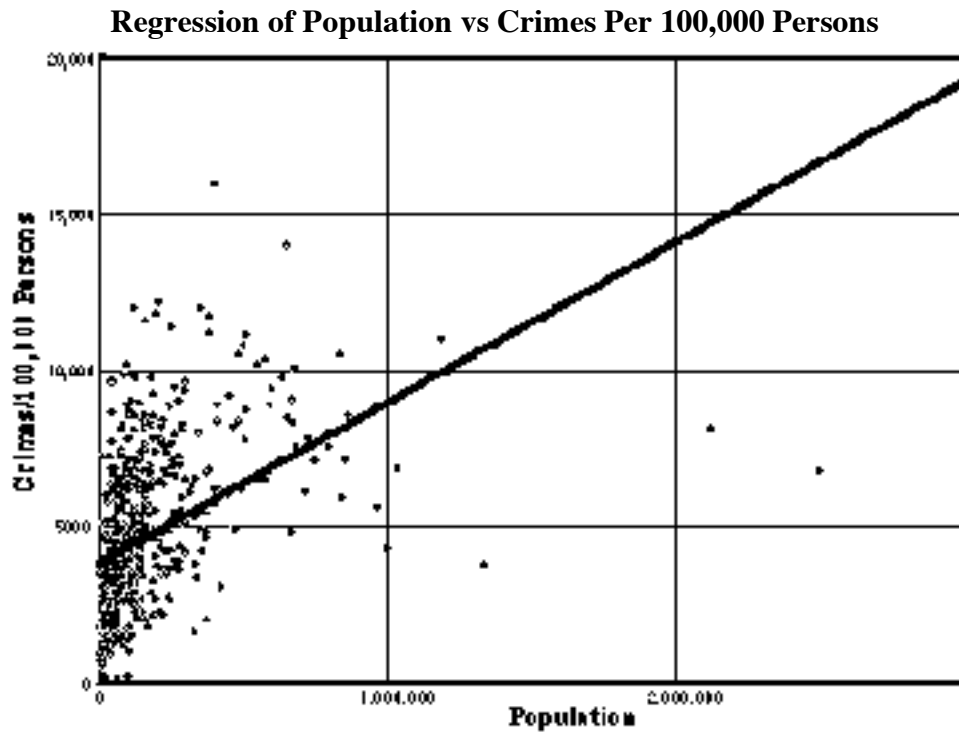
If the 593 counties had been a sample, the  $t$  test yields a value of 12.204. Consulting a table of  $t$ -statistic values indicates that a score of 1.96 would be expected to occur by chance only 5% of the time and 3.922 only .01% of the time. The value of 12.204 is far beyond that and so the null hypothesis could be rejected. This means that there really is a relationship between city size and crime rate. However, city size only accounts for 20% of the variations in crime rate between cities.

There are a number of assumptions made about the data in correlation analysis which are not always met. For example, the observations should be selected randomly, they should be measured on the interval or ratio scale and be normally distributed, and they should be independent of each other. The latter condition may be a particular problem in samples that are geographically near to one another, however, large sample sizes can mitigate many of these problems. The size of the geographic units may also play a part in the correlation score. Termed the modifiable areal unit problem, the size of the unit areas may affect the correlations of the paired variables. Thus one needs to express these statistical associations and conclusions in terms of the areal units actually used rather than make a general statement on association between variables.

## D. Regression

If the correlation between two variables is found to be significant and there is reason to suspect that one variable influences the other then one might decide to calculate a regression line for the two variables. In this example one might state that an increase in population results in an increase in the crime rate. Thus, the crime rate would be considered a dependent variable and the population size would be considered an independent variable. When plotting these variables, the dependent variable, crime, would be plotted on the  $y$ -axis and the independent variable would be plotted on the  $x$ -axis of a scattergram.

Regression expresses the relationship between the two variables as the equation for a line which best fits the scatter of points in a scattergram. The line minimizes the sum of the squared deviations of the dependent (y variable) from the line. From the equation one can estimate the value of y for a given value of x. Differences between the estimated and real y-axis values are residuals.



The equation for the above regression line is

$$\text{Crimes}/100\text{k} = 3897.35 + 0.005149 * \text{Pop}$$

The farther a given dot is from the regression line, the larger the residual. The residuals are of special interest because they represent exceptions to the general association expressed by the regression line. In the example of city size and crime rate, identify the cities represented by the largest eight to ten residuals as they appear to you on the scattergram.

Points lying far above the regression line represent cities which have much higher crime rates than are expected based on their population size; points lying far below the line represent cities with much lower than expected crime rates.

Since it is possible that quite different scatters of points could produce the same line, it is also helpful to calculate the *standard error of the estimate* which provides an indication of the scatter of the points about the line. This value can be useful for comparing different samples.

$$\text{SE of Est} = \text{SQRT} \left( \frac{\sum(Y - \hat{Y})^2}{N} \right)$$

N

For this crime example the standard error of the estimate is 2252.9

The reliability of the regression model also may be tested with analysis of variance. With the F statistic one can determine how much of the total y variability is due to the regression line and how much is due to the residuals. If a large portion of the variance comes from the equation and the independent variable, then the model provides a good prediction of y and a high value of F.

$$F = \frac{\frac{(\sum(Y-\sum Y/N))^2}{df}}{\frac{(\sum(Y-\sum Y/N))^2}{n - df - 1}}$$

Where df is the degrees of freedom.

For the crime example, the F statistic is 148.94. The null hypothesis would state that the regression model fails to predict the variation in y and could, by chance, generate a value of 3.86 (from a table of F statistics) 5% of the time. Thus the null hypothesis can be rejected.

## E. Exercises

Ex 8. Association between Variables